PRIVACY-PRESERVING BIOMETRIC PERSON IDENTIFICATION

Oleksiy Koval, Sviatoslav Voloshynovskiy and Thierry Pun

Department of Computer Science, University of Geneva
7, route de Drize, 1227, Geneva, Switzerland
phone: + (41) 22 379 01 56, + (41) 22 379 01 58, + (41) 22 379 01 53,
fax: + (41) 22 379 00 79,
email: Oleksiy.Koval@cui.unige.ch, svolos@cui.unige.ch, Thierry.Pun@cui.unige.ch
web: sip.unige.ch

ABSTRACT

In this paper we investigate biometric person identification. We model this process of person identification as multiple hypothesis testing and consider performance measures that can be attained in such a protocol in terms of exponents of average error probability. A special attention is paid to the privacy related issues. In particular, we consider performance/privacy trade-off due to the indirect identification that is based on data projection of an original data to a secret subspace. Finally, we approximate the obtained performance limits using properties of random projections. Finally, experimental simulations are used to exemplify our findings.

1. INTRODUCTION

Biometric person identification has become an unavoidable security feature serving as a basis for person identification and access control. Being a replacement of classical security tokens as passwords, pins, ID cards or passports, physical or behavioral features of humans have certain advantages. The major one concerns their inherent link with a user and thus infeasibility to be stolen, forgotten or lost. They are capable as well of providing sufficient accuracy of identification procedure in terms of producing an error during identity verification. Finally, biometrics are omnipresent, easy to acquire, sufficiently permanent and universal in the sense that are in possession of all people.

Despite the mentioned advantages, one should carefully address the question of biometric storage. In many protocols (like those that are used to grant the access to laptops, mobile phones, and portable hard drives) it is necessarily to store the original biometric template. Thus, via gaining the physical access to the device, the attacker can obtain the access to the mentioned original template that leads to various security and privacy compromised issues. Possible ways of such illegal actions are creating of faked biometrics, system impersonalization, or even illegal people tracking.

This security fraud was recently realized [1] and suggested to formulate the problem of biometric-based authentication as extraction of “common randomness” [2, 3] that is performed using the fundamentals of distributed source coding [4]. However, this approach assumes the availability of the original biometric template as a secret shared by distributed encoder and decoder. Thus, one can conclude that the realized security hole is not completely overpassed. At the same time, the use of hashes increases the probability of collusions.

Thus, the main goal of the paper consists in the analysis of biometric-based person identification from a joint performance-privacy/security perspective. We propose, instead of using original biometric templates for identity verification, to perform indirect identification based on their projection to a secure key-dependent transform domain. The role of the projecting operator is twofold: besides enhancing the security of the identification protocol, it is used for accomplishing computational complexity improvement via input data dimensionality reduction. Moreover, the analog of identification system was not considered.

The remaining part of the paper is organized in the following way. The problem under investigation is formulated in Section 2. Performance analysis of the identification protocol based on random projections is analysed in Section 3 versus direct identity verification. Section 4 contains experimental validation results. Finally, Section 5 concludes the paper and draws some future extensions of the obtained results.

Notations. We use capital letters to denote scalar random variables X, bold capital letters to denote vector random variables $\mathbf{X}$, corresponding small letters x and $\mathbf{x}$ to denote the realizations of scalar and vector random variables, respectively. The superscript N is used to denote length-N vectors $\mathbf{x} = \{x[1], x[2], \ldots, x[N]\}$ with *ith* element $x[i]$. We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to indicate that a random variable X is distributed according to $p_X(x)$. Calligraphic fonts $\mathscr{X}$ denote sets $X \in \mathscr{X}$ and $|\mathscr{X}|$ denotes a cardinality of set. Finally, $\mathbb{I}_N$ denotes a $N \times N$ identity matrix.

2. PROBLEM FORMULATION

A general diagram of identity verification based on secure biometrics and random projections is presented in Figure 1.

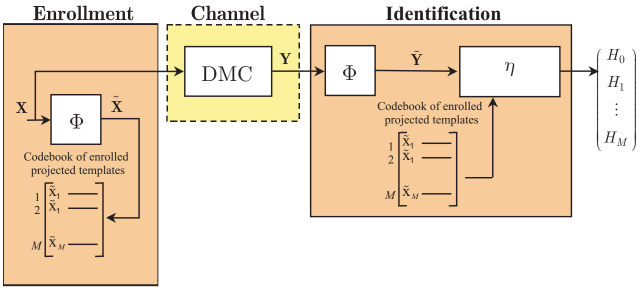


Figure 1: Person identification based on random projections: $\mathbf{\Phi}$ is a key-dependent random projection operator.

According to the presented set-up, the template identities are deduced during the enrollment stage based on the original template $\mathbf{X} \sim p(\mathbf{x})$ transformed to a secret key-dependent reduced dimensionality domain via applying a random projection operator Φ. These sequences $\mathbf{x}_i, i \in \{1,2,...,M\}$ are stored in the corresponding identification data base that is available at identification stage. The maximal amount of such indexes is limited by the maximum rate of reliable communications of the corresponding discrete memoryless channel (DMC) $p(y|x)$ that models the acquisition distortions. According to the channel coding theorem [5], $M \leq 2^{NI(X;Y)}$.

For the identity verification the template $\mathbf{Y}$ acquired from the output of $p(y|x)$ is projected to the mentioned secrete key-dependent transform domain $\tilde{\mathbf{Y}} = \Phi\mathbf{Y}$. Finally, a statistical test η is applied to decide which template out of $M = 2^{NI(X;Y)}$ alternatives is observed by the identity verification system.

As it was indicated in the introductory part of the paper, the main our goal is to evaluate the system performance in terms of identification accuracy and computational complexity impacted by the random projection. For this purpose, we model the problem of person identification as multiple hypothesis testing [6] and formulate it for the direct identification case where $\Phi = \mathbb{I}$. In particular, it is supposed that a source of biometric signals represented by a probability distribution $p(\mathbf{x})$ produces a length N vectors $\mathbf{X} \in \mathscr{X}^N$. The principle goal of the hypothesis test is to decide, which out of M possible cases is observed. We define an M-ary multimodal biometric person identification system as composed of the set $\mathscr{X}^N$ and a hypothesis test:

$$\eta : \mathscr{X}^N \rightarrow \{1,2,...,M\}, \tag{1}$$

where $\{1,2,...,M\}$ is a collection of possible alternatives.

The test is accomplished assuming the following prior distributions on respective hypotheses:

$$H_i : \mathbf{X} \sim p^i(\mathbf{x}) = \mathscr{N}(\mathbf{x}(i), \sigma_Z^2 \mathbb{I}_N), \tag{2}$$

where $i \in \{1,2,...,M\}$, $H_i \sim p^i(\mathbf{x}) = \prod_{j=1}^{N} p^i(x_j)$, and the acquisition channel is modeled as i.i.d. Gaussian with zero mean and variance σ_Z^2.

We measure the performance of the defined M-ary hypothesis testing using Bayesian average probability of error:

$$C(\eta) = \sum_i \pi_i \sum_{j \neq i} P_{ij}^E, \tag{3}$$

where η designates a selected decision rule, P_{ij}^E stays for the probability of falsely accepting hypothesis H_j in the case when H_i is in "true" and π_i is the prior probability of the hypothesis H_i.

We define the respective exponent of this average probability of error as follows:

$$\theta(\eta) = \lim_{N \to \infty} -\frac{1}{N} \log C(\eta). \tag{4}$$

Theoretical performance analysis of multiple hypothesis testing for the case of more than two hypotheses under the selected performance criterion of the average probability of error was found intractable [7, 8]. Instead of direct problem analysis, usually it is tackled exploiting some strategies that simplify the further analysis and considerations. One of such simplifications was firstly introduced in [8], where for the sake of analysis tractability the multiple hypothesis testing was replaced by a set of binary tests (Figure 3). It is known that in this case, the optimal decision rule follows the multiple maximum a posteriori (MMAP) strategy:

$$\eta^{MMAP} = \arg \max_{i \in \{1,2,...,M\}} \pi_i p^i(\mathbf{x}). \tag{5}$$

In the forgoing analysis we assume that prior hypothesis distribution, $\pi_i = \frac{1}{M}$, is uniform and MMAP rule will be simplified to multiple maximum likelihood (MML) decision rule:

$$\eta^{MML} = \arg \max_{i \in \{1,2,...,M\}} p^i(\mathbf{x}). \tag{6}$$

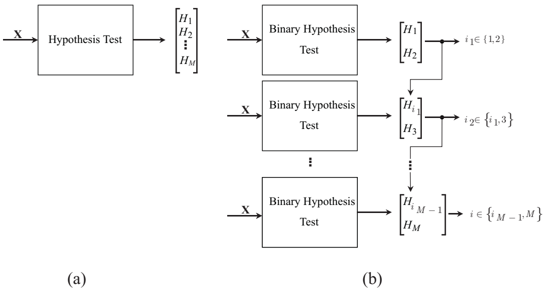


Figure 2: Multiple hypothesis testing (a) vs its approximation by a set of binary tests (b).

It was demonstrated [8] for the memoryless case that $\theta(\eta^{MMAP})$ is upper bounded as:

$$\theta(\eta^{MMAP}) \leq \theta(\eta^{MAP}) \leq D^*, \tag{7}$$

where $\theta(\eta^{MAP})$ designates a MAP detector for a binary hypothesis testing formulation and:

$$D^* = \min_{(m,n) \in \{1,2,...,M\}, m \neq n} D_s(p^m(x), p^n(x)), \tag{8}$$

where $D_s(p^m(x), p^n(x))$ denotes the Chernoff distance [5] between two distributions $p^m(x)$ and $p^n(x)$. This distance is defined in the following way:

$$D_s(p^m(x), p^n(x)) =$$

$$\max_{0 \leq s \leq 1} -\log \int_{\mathscr{X}} p^n(x) \left\{ \frac{p^m(x)}{p^n(x)} \right\}^s dx. \tag{9}$$

Thus one can conclude, that the performance of biometric person identification modeled as the M-ary hypothesis testing can be approximated by a set of binary tests and upper limited by the worst pairwise Chernoff distance between alternative hypotheses.

3. PERFORMANCE ANALYSIS

Direct identification. Performance analysis of the problem of biometric person identification formulated in the previous Section will be analysed assuming its Gaussian formulation and availability of the observation of the acquired genuine biometric template at verification. In fact we suppose that $p^i(\mathbf{x}) = \mathscr{N}(\mathbf{x}(i), \sigma_Z^2 \mathbb{I}_N)$. Such a formulation is justified by the assumption that equal portion of random distortions is added to an ideal biometric pattern during enrollment. Thus, one can rewrite (2) in the following form:

$$H_i : \mathbf{y} = \mathbf{x}(i) + \mathbf{z}, \tag{10}$$

where $i \in \{1,2,...,M\}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbb{I}_N)$.

For such a formulation of the problem under consideration and following the strategy of multiple hypothesis testing performance limitation by a binary test that considers the pair of hypotheses, *m* and *n* with the smallest Chernoff distance, the corresponding MML test (6) can be reconsidered as follows:

$$\eta^{ML} = \arg \max_{i \in \{m,n\}} p^i(\mathbf{x}). \tag{11}$$

Then, taking the logarithms of the likelihood functions, one obtains the rule that declares the hypothesis *m* being in force:

$$\log(p^m(\mathbf{x})) - \log(p^n(\mathbf{x})) \geq 0. \tag{12}$$

Therefore, one can rewrite (12) using sufficient statistics:

$$t(\mathbf{y}) = \mathbf{y}^T(\mathbf{x}(m) - \mathbf{x}(n)) - \frac{1}{2}(\varepsilon(m) - \varepsilon(n)), \tag{13}$$

where $\varepsilon(m) = \mathbf{x}(m)^T\mathbf{x}(m)$, $\varepsilon(n) = \mathbf{x}(n)^T\mathbf{x}(n)$ are the energies of vectors $\mathbf{x}(m)$ and $\mathbf{x}(n)$, respectively.

Then, the corresponding prior models on binary hypotheses can be expressed in terms of sufficient statistics as follows:

$$\begin{cases} H_n : & T(\mathbf{Y}) \sim \mathcal{N}(-\frac{1}{2}d^2, \sigma_Z^2 d^2), \\ H_m : & t(\mathbf{Y}) \sim \mathcal{N}(\frac{1}{2}d^2, \sigma_Z^2 d^2), \end{cases} \tag{14}$$

where $d = \|\mathbf{x}(m) - \mathbf{x}(n)\|^2$ is a square distance between original biometric patterns $\mathbf{x}(m)$ and $\mathbf{x}(n)$. Therefore, under the memoryless assumption and according to (9) the exponent of the average probability of error that one can expect to attain in biometric person identification modeled as multiple hypothesis testing is bounded by the corresponding Chernoff distance $D_s(p^m(x), p^n(x))$ that for the case of Gaussian statistics of hypothesis priors can be written as:

$$D_s(p^m(x), p^n(x)) = \frac{d^2}{8\sigma_Z^2}. \tag{15}$$

Furthermore,

$$\begin{aligned} d^2 &= \|\mathbf{x}(m) - \mathbf{x}(n)\|^2 \\ &= \|\mathbf{x}(m)\|^2 + \|\mathbf{x}(n)\|^2 - 2\mathbf{x}(m)^T\mathbf{x}(n) = 2\|\mathbf{x}\|^2 \end{aligned} \tag{16}$$

assuming that all biometric patterns have equal energy, i.e., $\|\mathbf{x}(m)\|^2 = \|\mathbf{x}(n)\|^2 = \|\mathbf{x}\|^2$, and all of them are pairwise orthogonal, i.e., $\mathbf{x}(m)^T\mathbf{x}(n) = 0$. Finally, one can reduce (15) as follows:

$$D_s(p^m(x), p^n(x)) = \frac{\|\mathbf{x}\|^2}{4\sigma_Z^2}. \tag{17}$$

Random projection-based identification. In this subsection of the paper we assume that genuine identification templates are not available for identification and identity verification is performed based on the data securely transformed to a certain secret domain via applying a secure random projection operator Φ:

$$\mathbf{y}' = \Phi\mathbf{y}. \tag{18}$$

Operator Φ besides security concerns serves to convert an original length-N vector to a vector of lower dimensionality $L, L \leq N$. An $L \times N$ operator Φ is a random matrix whose elements are generated from a certain density. In the scope of this paper we assume that the elements of Φ are generated from a zero mean Gaussian distribution with a variance $\frac{1}{N}$, i.e., $\mathcal{N}(0, \frac{1}{N})$.

It is possible to demonstrate that the corresponding prior probabilities of the worst case binary hypothesis test can be defined as follows:

$$\begin{cases} H_i' : & T'(\mathbf{Y}) \sim \mathcal{N}(-\frac{1}{2}d'^2, \sigma_Z^2 d'^2), \\ H_j' : & T'(\mathbf{Y}) \sim \mathcal{N}(\frac{1}{2}d'^2, \sigma_Z^2 d'^2), \end{cases} \tag{19}$$

where $d'^2 = (\mathbf{x}(m) - \mathbf{x}(n))^T\Phi^T(\Phi\Phi^T)^{-1}\Phi(\mathbf{x}(m) - \mathbf{x}(n))$ and Φ^T is the result of transposition of projection operator Φ.

Then the exponent of the average probability of error we are seeking for is defined:

$$D_s'(p^m(x), p^n(x)) = \frac{d'^2}{8\sigma_Z^2}. \tag{20}$$

In the case one assumes that this operator is an ortho-projector, i.e., $\Phi\Phi^T = \mathbb{I}_N$, the distance d' can be modified as follows: $d'^2 = (\mathbf{x}(m) - \mathbf{x}(n))^T\Phi^T\Phi(\mathbf{x}(m) - \mathbf{x}(n)) = \|\Phi(\mathbf{x}(m) - \mathbf{x}(n))\|^2$.

Attractiveness of projecting onto a randomized basis was first formulated by Johnson and Linderstrauss [9]. It was proved that in the case the projection is performed from a certain vector space onto a random subspace of sufficiently high dimension, the distance between the space elements will be approximately preserved. The following sandwich of inequalities is valid:

$$\begin{aligned} (1-\delta)\sqrt{\frac{L}{N}}\|\mathbf{x} - \mathbf{y}\| &\leq \|\Phi(\mathbf{x} - \mathbf{y})\| \\ &\leq (1+\delta)\sqrt{\frac{L}{N}}\|(\mathbf{x} - \mathbf{y})\| \end{aligned} \tag{21}$$

for a sufficiently small δ. Then, one can approximate (20) for the case of $\Phi\Phi^T = \mathbb{I}_N$ as presented below:

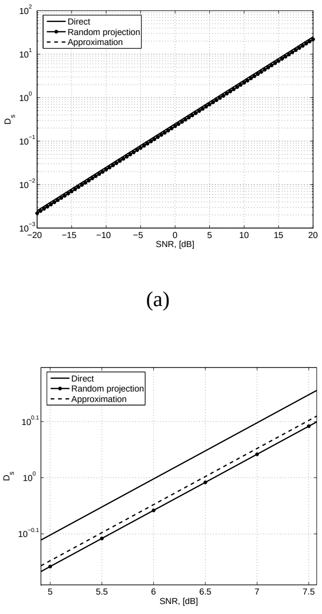
$$D_s'(p^m(x), p^n(x)) \approx \frac{L}{N}\frac{\|\mathbf{x}(m) - \mathbf{x}(n)\|^2}{8\sigma_Z^2}. \tag{22}$$

Finally, assuming that $\mathbf{x}(m), \mathbf{x}(n)$ are orthogonal sequences of equal energy, i.e., condition (16) is valid, one obtains:

$$D_s'(p^m(x), p^n(x)) \approx \frac{L}{N}\frac{\|\mathbf{x}\|^2}{4\sigma_Z^2}. \tag{23}$$

4. EXPERIMENTAL VALIDATION

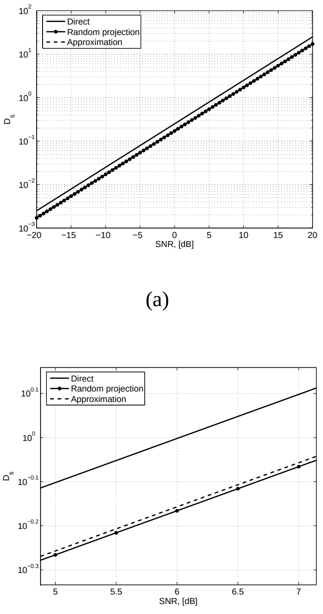
The main goal of this Section is threefold. We will try to investigate the behavior of exponent of the average probability of error as a function of operational signal-to-noise ratio (*SNR*) defined as $SNR = 10\log_{10}\left(\frac{\|\mathbf{x}\|^2}{\sigma_Z^2}\right)$ for the direct identification case (17). Secondly, we would like to estimate the loss in performance in the case the identification is performed based on random projection (20). Finally, we would



(a)

(b)

Figure 3: Behavior of exponent of average probability of error for direct identification, random projection-based identification and approximation of random projection-based identification, $\frac{L}{N} = 0.9$: (a) full-size plot and (b) magnified fragment.



(a)

(b)

Figure 4: Behavior of exponent of average probability of error for direct identification, random projection-based identification and approximation of random projection-based identification, $\frac{L}{N} = 0.7$: (a) full-size plot and (b) magnified fragment.

like to see the accuracy of performance approximation (20) using Johnson-Linderstrauss result (23). For this purpose we run a set of experimental simulations according to the following setup: the length of all vectors N was selected to be equal 1024, the operational SNR range was fixed to the interval $[-20; 20]$ dB. The templates $\mathbf{x}$ were generated from a zero-mean, unit variance Gaussian distribution, while the elements of random projector operator Φ are i.i.d. zero-mean Gaussian with the variance $\frac{1}{N}$. The dimensionality reduction ratio $\frac{L}{N}$ was set to $[0.9; 0.8; 0.7; 0.6; 0.5]$. Since the obtained results revealed similar behavior of error exponent of the average probability of error for all selected $\frac{L}{N}$, only the cases of $\frac{L}{N} = [0.9; 0.7; 0.5]$ are presented in the paper (Figure 3-5). The obtained experimental result allows to conclude that in multiple hypothesis testing biometric person identification problem formulation under the stationary acquisition conditions a certain loss in performance of random projection based identification versus a direct one is observed in terms of exponent of the average probability of error due to the impact of the projection operator. These results also confirm a high accuracy of Johnson-Linderstrauss approximation of performance of the random projection-based identification.

5. CONCLUSIONS AND FUTURE RESEARCH PERSPECTIVES

In this paper we considered the problem of privacy preserving biometric person identification from performance perspective. We propose to protect biometric templates used for person identity verification projecting them onto a secret randomize subset. We conducted performance analysis of random projection-based identification versus a direct one in terms of error exponent of the average probability of error. In the former case we were able to provide performance limits approximation using properties of random projections. By means of computer simulations we were capable of demonstrating a certain performance loss versus direct identity verification that is linearly proportional to the dimensionality reduction ratio. Finally, a high accuracy of Johnson-Linderstrauss approximation of performance of projection-based identification was demonstrated.

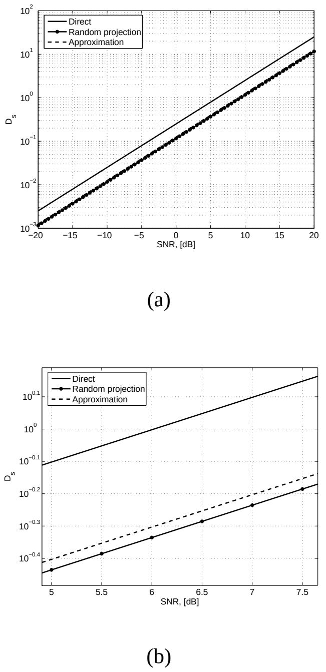
In our future research we will extend the performed analysis to the case of multibiometric person identification that is known to achieve even more accurate identification results.

6. ACKNOWLEDGMENTS

This paper was partially supported by SNF PB grant 114613 and SNF 200021–111643, by the EU projects IST-2002-507932-ECRYPT, FP6-507609-SIMILAR and Swiss IM2 project.

REFERENCES

[1] E. Martinian, S. Yekhanin, and J. Yedidia, “Secure biometrics via syndromes,” in *43rd Annual Allerton Conference on Communications, Control, and Computing*, Monticello, IL, USA, October 2005.



(a)

(b)

Figure 5: Behavior of exponent of average probability of error for direct identification, random projection-based identification and approximation of random projection-based identification $\frac{L}{N} = 0.5$: (a) full-size plot and (b) magnified fragment.

[2] U. Maurer, “Secret key agreement by public discussion from common information,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 733–742, 1993.

[3] R. Ahlswede and I. Csiszar, “Common randomness in information theory and cryptography - part i: secret sharing,” *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1121–1132, 1993.

[4] D. Slepian and J. Wolf, “Noiseless encoding of correlated information sourcea,” *IEEE Trans. Information Theory*, vol. 19, pp. 471–480, July 1973.

[5] T. Cover and J. Thomas, *Elements of Information Theory.* Wiley and Sons, New York, 1991.

[6] O. Koval, S. Voloshynovskiy, and T. Pun, “Analysis of multimodal binary detection systems based on dependent/independent modalities,” in *IEEE International Workshop on Multimedia Signal Processing (MMSP-07)*, Chania, Crete, Greece, October 2007.

[7] H. van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York, Wiley, 1968.

[8] C.C.Leang and D.H.Johnson, “On the asymptotics of m-hypothesis bayesian detection,” *IEEE Trans. on Information Theory*, vol. 43, no. 1, pp. 280–282, January 1997.

[9] W. B. Johnson and J. Lindenstrauss, “Extensions of lipschitz mapping into hilbert space,” *Contemporary Mathematics*, no. 26, pp. 189–206, 1984.